A WEB OF DATA ANALYTICS SERVICES

Thomas Huang

Jet Propulsion Laboratory, California Institute of Technology

4800 Oak Grove Drive, Pasadena, CA 91109, U.S.A.

ABSTRACT

Cloud Computing has become the ubiquitous approach to our Big Data challenge. However, one will quickly discover that moving (a.k.a. forklifting) existing on-premise data analytics solutions to the Cloud doesn't always translate to costing saving and performance boost. The Cloud's elasticity, its availability, and its wide selection of computing options and selections of costing models making Cloud an attractive environment to tackle our Big Data challenge. The fact is Cloud, on its own, is not the silver bullet to our daunting challenge need for analyze and derive scientific inferences through vast collections of multi-sensor measurements. We would like to have all scientific data in one easy to access environment, but getting the world of scientific data in one analytic system is immensely difficult to achieve. This paper describes the data analytics web architecture NASA is developing by infusing instances of Integrated Data Analytics systems next to the data. The goal is to minimize unnecessary data movement through collection of data access and analytics webservices for researchers to interact with and analyze measurements without have to download data to their local computer. These services are RESTful and provisioned by the data centers with the help from subject matter and science experts. These services encapsulate the physical computing infrastructure, which could local computing cluster, on-premise or public Cloud environment.

Index Terms— Big Data, Distributed Analytics, Parallel Analytics, Cloud Computing, Ocean Science, OceanWorks, Apache SDAP, Apache NEXUS, CEOS, PO.DAAC, NASA

1. INTRODUCTION

Climate change is a defining issue of our time. It is touching on direct human and societal impacts. With increasing global temperature warming of the ocean and melting ice sheets and glaciers, the impacts can be observed from our coastline, and may involve drastic changes to marine ecosystems. While there is no lack of information and publications on climate change impacts on aspects such as sea level rise, floods, droughts, and hurricanes, understanding of ecosystem level impacts on and effects on flood security is critically important yet very poorly understood. Adding to the science and data integration challenges that understanding these impacts poses is the complexity of broader public and policy-maker engagement as stakeholders and fundamental determinants of future outcomes.

While much of the satellite observations from various disciplines are accessible from different data centers, the solution for analyzing decades of measurements and coordinating measurements collected from various instruments for time series analysis is both difficult and critical. Climate research is a big data problem that involves high data volume, measurements collected by various sources, methods for on-the-fly extraction and reduction to keep up with the speed and data volume, and the ability to address uncertainties from data collections, processing, and analysis.

For decades scientists have been relying on a common process flow, which includes scrape FTP sites, download data files to their local computing environment, and developing algorithms to analyze the downloaded data. Data center are only chartered to distribute file products. In this age of big data, our climate research community recognizes the traditional analytic workflow is unsustainable. While data centers do provide some tools for reduction, such as data subsetters, the size of the subsetted data may still be too large to download. A more efficient approach is to have large analytic solutions right next to the data holdings to eliminate data movement. With affordable Infrastructure as a Service (IaaS) of commercial Cloud and semantic web, we are still seeing much of the informatics community is in the business of building one-off, stovepipe tools. Users are finding themselves working with different disjoint tools and having to manually translate between different data formats and nomenclatures, often data have to be transformed into different representations to satisfy different tools requirements.

We need a web of Integrated Data Analytics systems that shares common taxonomy and provides common webservice API for access and analysis that allows the service providers to scale-up or scale-down the computing according to the requirements and user needs. The users of these services shouldn't have to be concerned about the physical computing and internal data management architecture. More importantly, these services share common taxonomy and nomenclature to enable federated analysis of different measurements.

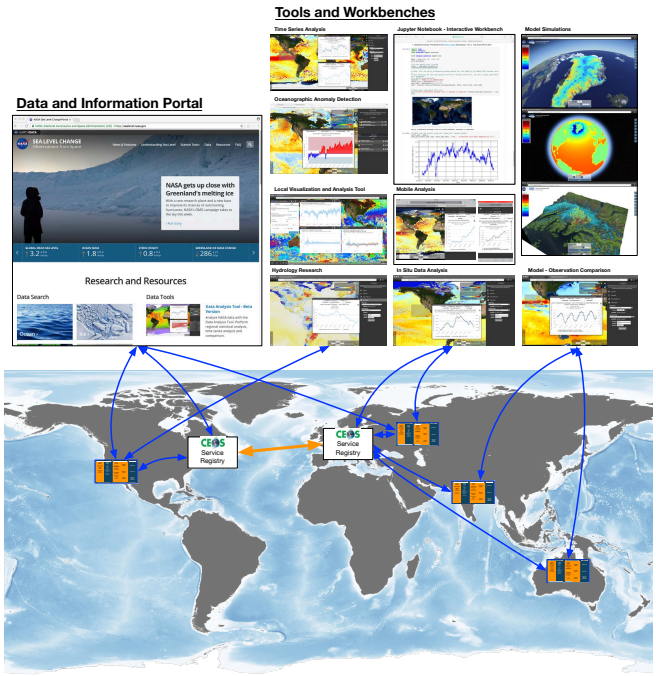

FIGURE 1: DISTRIBUTED ANALYTICS ARCHITECTURE

2. DISTRIBUTED ARCHITECTURE

The Committee on Earth Observation Satellites (CEOS) Ocean Variables Enabling Research and Applications for GEOS (COVERAGE) initiative [8] is an international initiative that seeks to provide improved access to multi-agency ocean remote sensing that are better integrated with in-situ and biological observations, in support of oceanographic and decision support applications for societal benefit. While it would be ideal to have all data in one place, such as a common Cloud computing environment, such solution is unsustainable due to various factors including international policies between agencies and security requirements, access to subject matter or domain experts, and the overall cost for managing and providing open access to exabyte (EB) of data in one place. COVERAGE has taken on a distributed analytic architectural approach [1] where each data provider or agency can standup their own Integrated Data Analytics Platform for the data they manage. The services share common API, taxonomy, and metadata model. All analyzed results are packaged in JSON documents. This architectural approach reduces the need for unnecessary massive data movement between services and the client application will only have to develop logics to process the result JSON responses. CEOS Service Registry can be established according to the continents and/or agency alliance to serve as the data and services lookup and discovery access point.

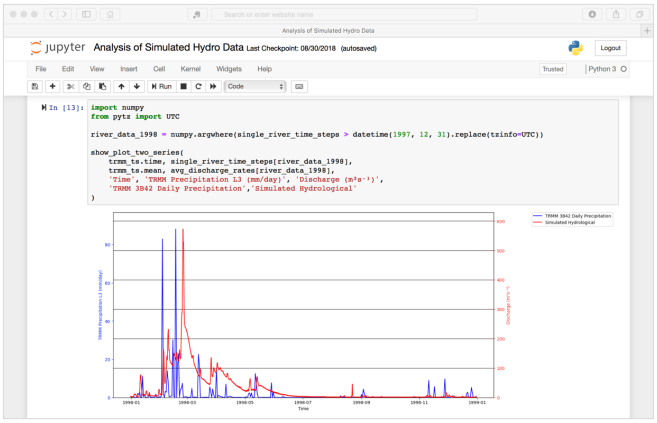

FIGURE 2: PLOTTING TIME SERIES BETWEEN RIVER AND TRMM PRECIPTATION MEASRUEMENTS

Clients of COVERAGE include

- *Data portals* for data and climatological events discovery that link to relevant data, analytics services and published results.
- *GIS-based domain-specific data tools* that is tailored to specific science investigation and/or community. Examples of such tools include
    - NASA Sea Level Change Portal's Data Analysis Tool (http://sealevel.nasa.gov/data-analysis-tool/) [3] is an advanced data visualization and analysis tool for sea level rise research
    - The GRACE Data Analysis Tool (https://grace.jpl.nasa.gov/data-analysis-tool/) is an advanced analysis tool specifically for the GRACE data.
    - The NASA Physical Oceanography Distributed Active Archive Data Center (PO.DAAC)'s State of the Ocean Tool (https://podaac-tools.jpl.nasa.gov/soto/) is a web-based tool for physical oceanography data.
- *Domain-specific applications* which could ranges from simple scripts to advanced GIS-based programs iin any programming languages (e.g. Python, Java, MATLAB, IDL, C/C++, etc.) to orchestrate search results and analytic operations.
- *Interactive workbench*, such as the popular Jupyter Notebook (https://jupyter.org), for researchers to interact with these services to create recipes to share with other researchers. Fig. 2 is an example of an interactive workbench demonstrated at the 2018 CEOS SIT Technical Workshop at Darmstadt, Germany [1]. The demo generated coordinated time-series between river gauges and perception data from the Tropical Rainfall Measurement Mission (TRMM). The river time series was computed by an analytic service at the NASA JPL and the TRMM time series was produced by the analytic service hosted under the Amazon Web Services (AWS). This demo involved no data movement. The Jupyter

Notebook was running on a typical laptop computer connected to the internet over WIFI. The demo shows the spike on river runoff after abnormal rate of rainfalls around February 1998 in the county of Los Angeles.

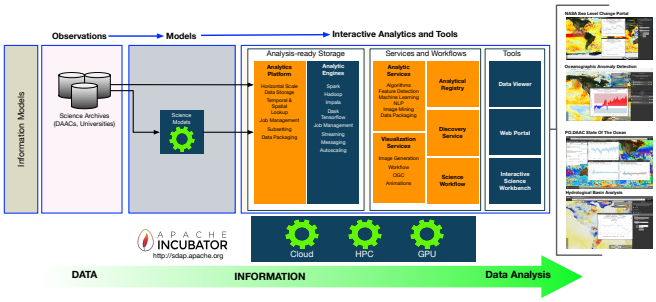

FIGURE 3: ARCHITECTURE FOR AN INTEGRATED DATA ANALYTICS PLATFORM

3. INTEGRATED DATA ANALYTICS PLATFORM

An Integrated Data Analytics Platform is an architectural concept to encapsulate the scalable computational and data infrastructures and to harmonize data, tools and computation resources to enable scientific investigations. The goal is to create a webservice platform for researchers and tools developers to discover, interact and analysis massive amount of related data without having to move data between systems over the internet. This platform must tackle both storage and software architecture together in order to fully leverage of its operating environment, such as the elastic cloud, without tying the users of the platform to a specific cloud provider and/or a specific underlying technology.

The Apache Science Data Analytics Platform (SDAP) (https://sdap.apache.org) is an open source implementation of an Integrated Data Analytics Platform. The technology is the backend for the NASA's Sea Level Change Portal, NASA's GRACE science portal, and the core for the NASA's Advanced Information Systems Technology (AIST) OceanWorks technology, which will be the analytics solution for the NASA's Physical Oceanography Distributed Active Archive (https://podaac.jpl.nasa.gov) for the ocean science community.

Fig 3 illustrates the architecture of an Integrated Data Analytics Platform [2]. Rather than aiming for creating a killer scientific application, the goal is to create a service platform to enable suite of scientific applications and systems. The platform can be divided into three tiers

1. *Tools and applications* – these are the clients of the platform. Their only binding to the platform is through RESTful APIs with all responds packaged in JSON documents. These clients can be implemented in any web-enabled programming languages, that is, able to make HTTP(S) calls and able to parse simple text response in JSON format. These clients have no knowledge of the physical hardware infrastructure and how the actual data is being stored.
2. *Services and Workflow* – these are the implementation of the data access and analytics webservices. They are the clients of the Analysis-Ready Storage tier. These services and workflow are built to leverage the parallel GIS-based data query and retrieval services provided by the Analysis-Ready Storage tier. It is a parallel analytic environment. The SDAP analytics services are implemented using Apache Spark for fast, in-memory MapReduce statistical analysis operations. It has no knowledge of how the data is physically stored and how the spatial indexes are being maintained. These services include area-averaged time series, climatological map, etc. The Workflow are for automated processing such as generation of climatology and large on-demand services.
3. *Analysis-Ready Storage* – it is more than a collection of disks and folders. The platform is designed for horizontal scaling, that is, to enable parallel fetching and apply parallel analytics. This tier harmonizes different satellite observation data and its metadata to create a unified representation of information to simplify the development of analytic webservices and workflow systems. It is also equipped with its own workflow system to automate the discovery, transformation, and ingestion of various new observational and model data from different data providers.

The deployment of such big data analytics solution is no small task if done manually. As a horizontal-scale solution, depending on the volume and the kind of analysis, it involves orchestration of large number of compute nodes. Container deployment technology, such as Kubernetes and Docker, has matured over the years. SDAP packages all of its components and services into a collection of Docker containers where the deployment can be automated using Continuous Integration (CI) tool such as Jenkins or Atlassian Bamboo.

3.1. NASA's OceanWorks project and the Apache Science Data Analytics Platform (SDAP)

OceanWorks is an NASA Advanced Information Systems Technology (AIST) project to establish an Integrated Data Analytics Platform at the NASA PO.DAAC for big ocean science. It focuses on technology integration, advancement and maturity by bringing together several previous NASA-funded analytics projects as an effort to deliver a production-ready data science platform for the ocean science community. OceanWorks is a key part of PO.DAAC's solution for analyzing 23PB of NASA's upcoming Surface Water Ocean Topography (SWOT) mission where its data will be hosted on the Cloud. Recognizing the building blocks of OceanWorks can support multi-disciplinary Earth Science, the OceanWorks project collaborates with the Apache Software Foundation and established the Apache Science Data Analytics Platform (SDAP) (https://sdap.apache.org). The goal is to establish a community-driven and supported GIS-based big data analytics platform. The components of SDAP includes:

- NEXUS: the big data analytics engine. See the following subsection.
- Extensible Data Gateway Environment (EDGE) [4]: a GIS-based OpenSearch and metadata translation integration service for fast geospatial lookup of data and translate metadata into various standards includes ISO-19115, DIF, UMM-C and UMM-G, etc.
- OceanXtremes [9]: a big data analytics solution for anomaly detection that enables to perform on-the-fly computation of daily difference by comparing observation against the climatology and provides tools for scientists to register anomalies and publish them using RSS feed.
- Distributed Oceanographic Matchup Service (DOMS) [7]: a big data analytics solution to perform on-the-fly matchup of in-situ measurements against satellite observation. To date, the in-situ data include SPURS I/II from JPL, SAMOS from the Center for Atmospheric Prediction Studies (COAPS) at Florida State University, and ICOADS from the National Center for Atmospheric Research (NCAR).
- Data relevancy [10] and event search: the data relevancy engine is a machine learning based technology to continuously analyze web search logs to dynamically rank the relevant datasets. The goal is to have the most relevant datasets listed in the beginning of the search results. The event search solution is to create relevant search respond that is encoded with space and time information. If a user searches for a specific hurricane, the responding datasets include URLs for the users to directly visualize and analyze the relevant data for a specific time period and location.

The Apache SDAP is currently under Apache Incubation process. It is in active development and infusion into various domain-specific environments.

3.2. Big data analytics engine

NEXUS (Fig. 4) is an emerging data-intensive analytics framework. It takes a different approach on handling file-based observational temporal, geospatial artifacts in order to fully leveraging existing horizontal-scaling technologies like MapReduce and the elastic cloud environment. NEXUS breaks the original data file into tiles and stores tiled data in cloud-scaled databases with an added high-performance spatial lookup service. NEXUS provides the bridge between science data and horizontal-scaling data analysis. This platform simplifies development of big data analysis solutions by bridging the gap between files and MapReduce solutions.

In addition to delivering the typical analytics services such as area-averaged time series and coordination map, NEXUS is also the base analytic framework for OceanXtremes and DOMS.

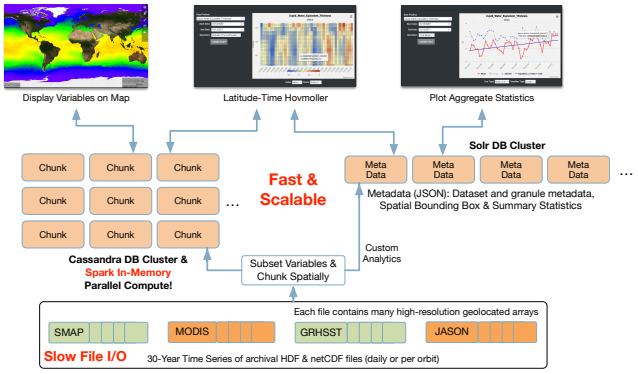

FIGURE 4: NEXUS' TWO-DATABASE ARCHITECTURE

NEXUS is designed to be adaptable to different deployment environments. It supports on-premise computing cluster and private/public cloud (such as AWS). It uses Apache Solr as its spatial registry for data tiles, metadata and pre-computed tile statistics. For data tile management, NEXUS supports fast, cloud-based NoSQL databases like Apache Cassandra and ScyllaDB, and it also supports storing tiles in an object store like AWS S3. For data ingestion, it uses serverless architecture when operate on the AWS and uses an ingestion cluster when operate on local hardware. The goal is to create a GIS-based analytics framework that is flexible to the project needs. Since this is a webservice-based solution, the internal infrastructure is hidden from the users of this framework.

3.3. Performance

NEXUS is still evolving as the community continuously finding new ways to improve its architecture and performance. A recent benchmark was gathered to analyze 16 years of MODIS TERRA Aerosol Optical Depth 550 nm (Dark Target) (MOD08_D3v6) [5][6] on a point-based, regional, and global scale. The analysis involves subsetting 5790 daily files (2.9GB) and apply analysis on the subsetted data. Performance numbers were gathers between NASA'G GIOVANNI, AWS EMR, and NEXUS. NASA's GIOVANNI is a popular web-based data analysis tool, that is built around file-based analysis. Fig. 5 shows NEXUS outperforms the traditional analysis method by hundreds of times. What usually takes nearly 30 minutes to compute, it only took NEXUS less than 2-second to compute.

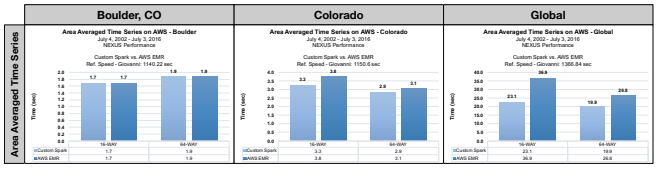

FIGURE 5: NEXUS PERFORMANCE

4. CONCLUSION

In the age of Big Data, we look to the Cloud as the solution to our challenge. We should consider Cloud as an instrument to our solution. In order to tackle our big data challenge and to deliver high performance analytic capacities to our climate researchers, we need to start with a scalable architecture. Our goal is to have our computing close to the data and deliver services for users to work with the data without the need of data download. The idea of Distributed Data Analytics relies on federated instances of Integrated Data Analytics systems. The demonstration and performance figures presented here have proven the importance of having a community-driven open source architecture for big data analytics in order to deliver end-to-end data management and horizontal-scale analytic services, which eliminates the need for massive data download and expensive hardware procurement for a domain-specific science investigation. The NASA OceanWorks will be infused into PO.DAAC to introduce on-the-fly capabilities to PO.DAAC's SOTO tool this year. Apache SDAP is expected to graduate from the Incubator this year as well.

5. ACKNOWLEDGEMENT

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

6. REFERENCES

[1] Huang, T. *Architecture for Distributed Earth Science Data Analysis*. In proceedings of the 2018 CEOS SIT Technical Workshop, Darmstadt, Germany, 2018.

[2] Huang, T., E.M. Armstrong, F.R. Greguska, J. Jacob, N. Quach, L. McGibbney, V. Tsontos, B. Wilson, S. Smith, M.A. Bourassa, J. Elya, S.J. Worley, T. Cram, and Z. Ji. *High Performance Open-Source Big Ocean Science Platform*. In proceedings of the 2018 Ocean Science Meeting, Portland, OR, 2018.

[3] Huang, T. *NASA Sea Level Change Portal – It is not just another portal site*. In proceedings of the 2017 American Geophysical Union Fall Meeting, New Orleans, LA, 2017.

[4] Huang, T., E.M. Armstrong, and N. Quach. *Metadata-Centric Discovery Service*. In proceedings of the 2012 Federation of Earth Science Information Partners (ESIP) Summer Meeting, Madison, WI, 2012.

[5] Jacob, J.C., F. R. Greguska, T. Huang, N. Quach, and B. D. Wilson. *Design Patterns to Achieve 300x Speedup for Oceanographic Analytics in the Cloud*. In proceedings of the 2017 American Geophysical Union Fall Meeting, New Orleans, LA, 2017.

[6] Lynnes, C., M. M. Little, T. Huang, J. C. Jacob, C. Yang, and K. Kuo. *Benchmark Comparison of Cloud Analytics Methods Applied to Earth Observations*. In proceedings of the 2016 American Geophysical Union Fall Meeting, San Francisco, CA., 2016.

[7] Smith, S.R., J. Elya, M.A. Bourassa, T. Huang, V. Tsontos, B. Holt, N. Quach, K. Gill, F. Greguska, S. Worley, and Z. Ji. The Distributed Oceanographic Match-Up Service. In proceedings of the 2018 Federation of Earth Science Information Partners (ESIP) Winter Meeting, Bethesda, MD, 2018.

[8] Vazuez, J., V. Tsontos, and E. Lindstrom. CEOS Ocean Variables Enabling Research & Applications for GEOS. In proceedings of the 19th International GHRSST Science Team Meeting (GHRSST XIX), Darmstadt, Germany, 2018.

[9] Wilson, B., E.M. Armstrong, T. Chin, K. Gill, F. Greguska, T. Huang, J. Jacob, and N. Quach. OceanXtremes: Scalable Anomaly Detection in Oceanographic Time-Series. In proceedings of the 2106 American Geophysical Union Fall Meeting, San Francisco, CA, 2016.

[10] Yang, C., E.M. Armstrong, M. Bambacus, K. Clarke, M. Cole, D. Duffy, S. Graves, W. Guan, Y. Jiang, K. Keiser, T. Huang, E. Law, Y. Li, Q. Liu, M. Little, D. Moroni, H. Qin, M. Rice, J. Schnase, D. Sherman, M. Xu, and M. Yu. *Big Data Platform* for Storing, Accessing, Mining and Learning Geospatial Data. In proceeding of 2017 American Geophysical Union Fall Meeting, New Orleans, LA, 2017.